# Know your data!

# Assumption testing and outlier identification

Ben Jann

ETH Zürich (Sociology)

e-mail: jann@soz.gess.ethz.ch

[Know Your Data! Assumption Testing and Outlier Identification Considering the Analysis of Reputation Effects in Internet Auctions as Example]

# Later entrants versus early birds:

[Does the Market Pay Off?, Wu and Xie 2003, ASR]

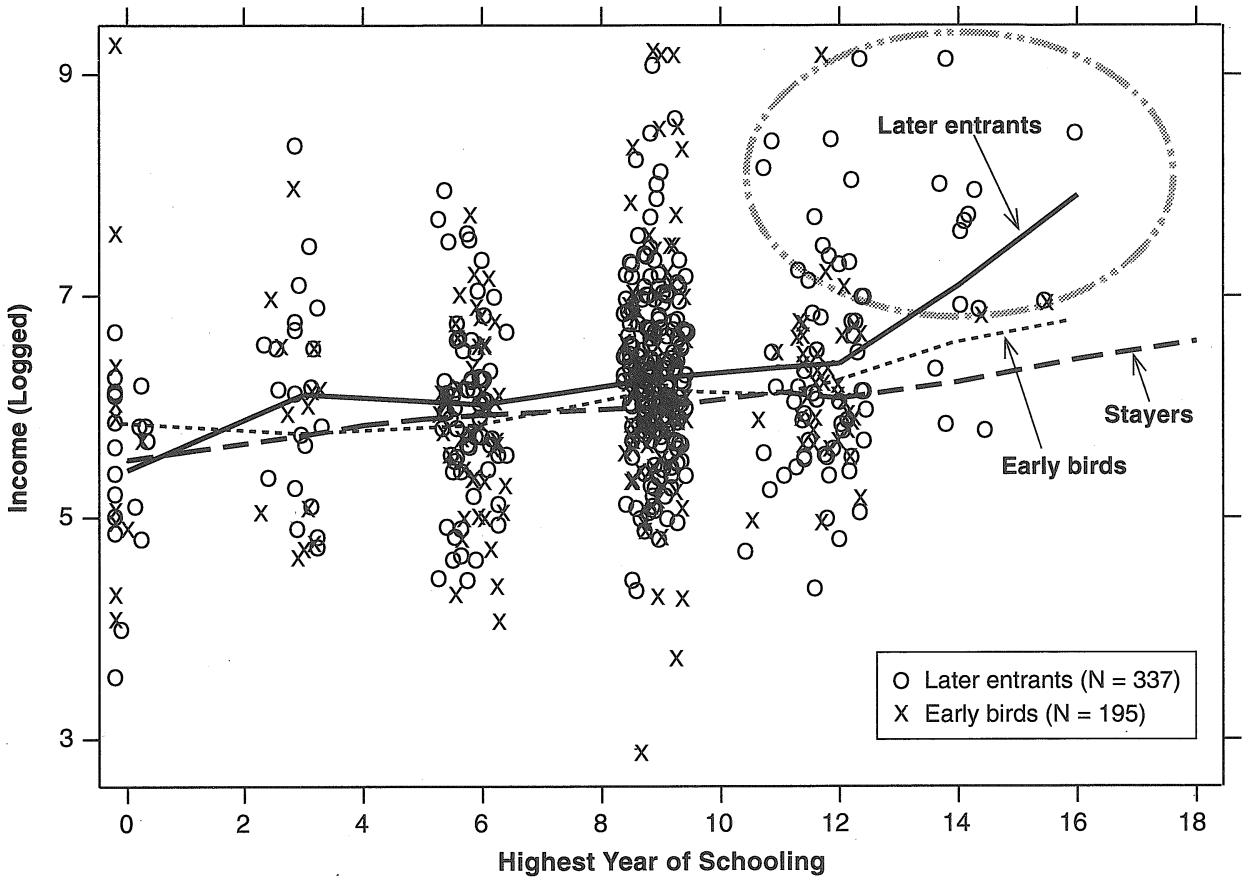


**Table 4. OLS Coefficients from the Multiple Linear Regression of Monthly Earnings on Selected Independent Variables, Urban China, 1996: Three-Worker-Type Analysis**

| | Restrictive Measure | | Broad Measure | | Comprehensive Measure | |
|---|---|---|---|---|---|---|
| Variable | Model 4a | Model 5a | Model 4b | Model 5b | Model 4c | Model 5c |
| Education (years of schooling) | .049*** (.009) | .045*** (.010) | .053*** (.008) | .045*** (.009) | .057*** (.007) | .047*** (.006) |
| Experience | .010** (.005) | .009* (.005) | .010* (.004) | .008 (.005) | .014*** (.004) | .013*** (.004) |
| (Experience)$^2$ × 1,000 | −.153* (.074) | −.144* (.070) | −.160* (.074) | −.099 (.070) | −.203** (.070) | -0.167** (.060) |
| Party member (yes = 1) | .121** (.038) | .126** (.037) | .138*** (.035) | .145*** (.035) | .142*** (.037) | .149*** (.037) |
| Sex (male = 1) | .218*** (.040) | .213*** (.040) | .220*** (.038) | .210*** (.039) | .225*** (.038) | .216*** (.038) |
| Later entrants[a] | .312* (.144) | −.732 (.370) | .238*** (.068) | −.263 (.193) | .313*** (.071) | −.175 (.182) |
| Early birds | .553 (.439) | .130 (.602) | .184 (.230) | −.124 (.266) | .151 (.206) | −.067 (.249) |
| Later entrants × Education | — | .122* (.047) | — | .060* (.022) | — | .056** (.019) |
| Early birds × Education | — | .051 (.103) | — | .037 (.031) | — | .025 (.024) |
| Constant | 5.305*** (.156) | 5.348*** (.165) | 5.238*** (.140) | 5.333*** (.153) | 5.124*** (.106) | 5.230*** (.116) |
| Number of cases | 2,072 | | 2,061 | | 2,060 | |
| $R^2$ | .117 | .129 | .114 | .123 | .127 | .136 |

*Notes*: Numbers in parentheses are standard errors adjusted for clustering on counties. Data are weighted.

[a] "Stayers" is the reference category; market losers are omitted from the analysis because of the small number of cases (N = 19).

*$p < .05$ **$p < .01$ ***$p < .001$ (two-tailed tests)

Our findings suggest that the commonly observed higher earnings and higher returns to education in the market sector compared with the state sector in China are due entirely to the earnings outcomes of later entrants. Early market entrants resemble work-

[Wu and Xie 2003: 438]

tural effect, rather than the second type. That is, the higher return to education is not caused by the market per se, but is associated with the characteristics of workers in the market sector.

[Wu and Xie 2003: 439]

## Review A:

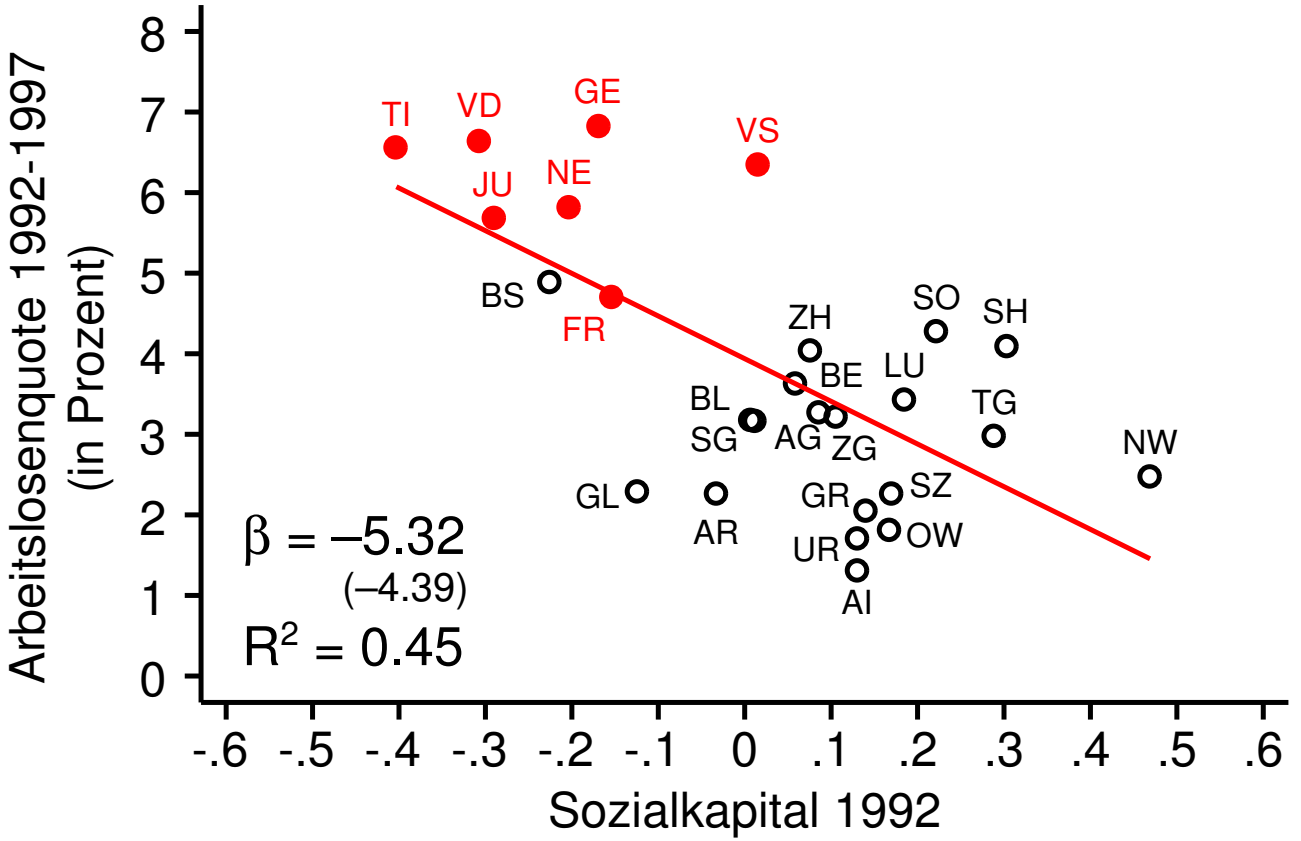
So, I conclude that the authors of this comment have identified a weakness of the Wu and Xie paper, and they should be congratulated for reading the article carefully.

But . . .

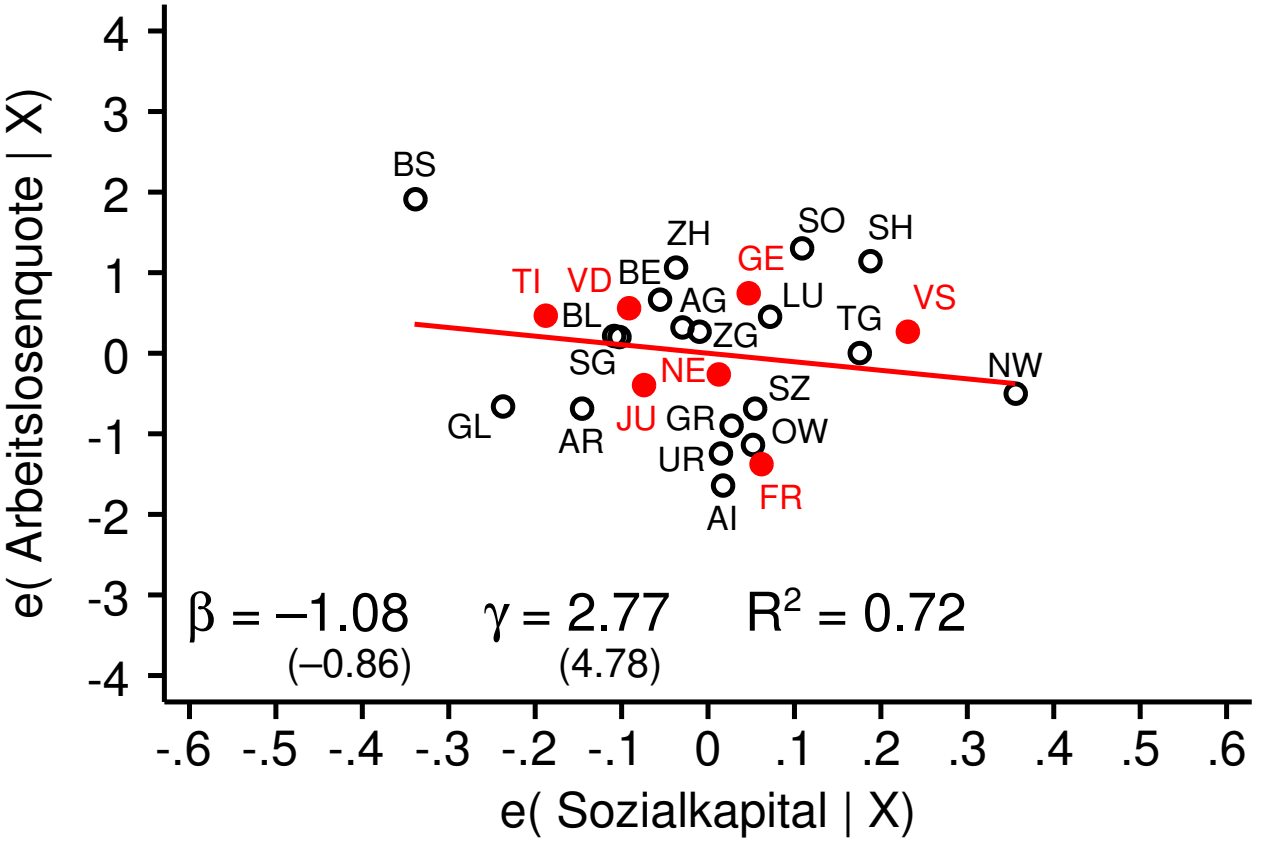
This sort of criticism could be leveled at, I suspect, 80% of all quantitatively-focused articles in ASR (and probably 95% of all quantitatively-focused articles in sociology). It is therefore quite embarrassing that Yu Xie, the recent Chair of the methodology section, would sign off on such a naïve analysis, but . . . then again . . . this seems to be the sort of simplistic analysis that gets one by a sociologist reviewer these days. That is, for years the highest payoff publication strategy has been: Put forward a very simple hypothesis about one coefficient being significant and a second simple hypothesis about another coefficient being significant. Then, run a regression model and proclaim one of the two hypotheses rejected.

# Social capital and unemployment:

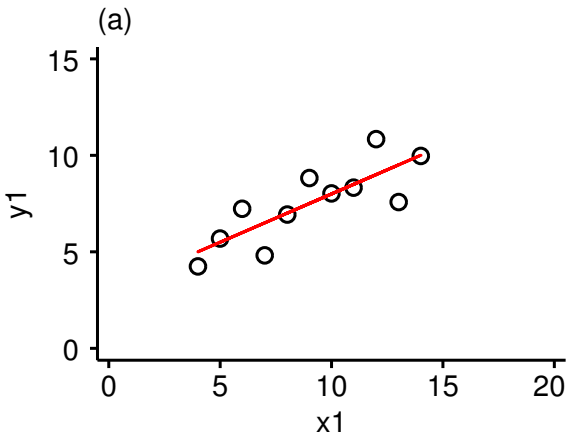
[Soziales Kapital und Arbeitslosigkeit, Freitag 2000, ZfS]

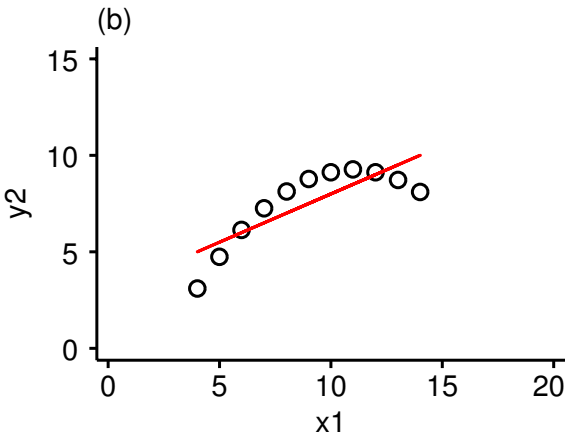


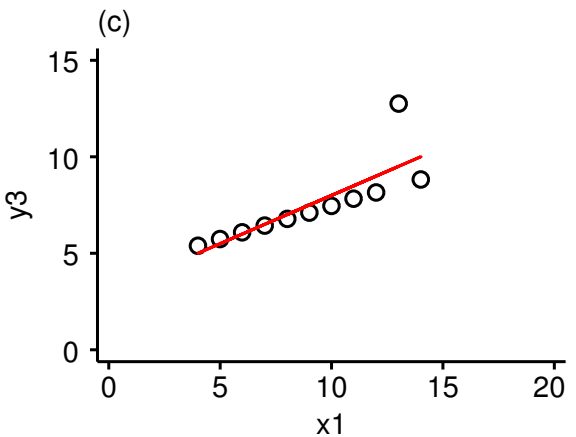
# Partial regression plot (added variable plot)

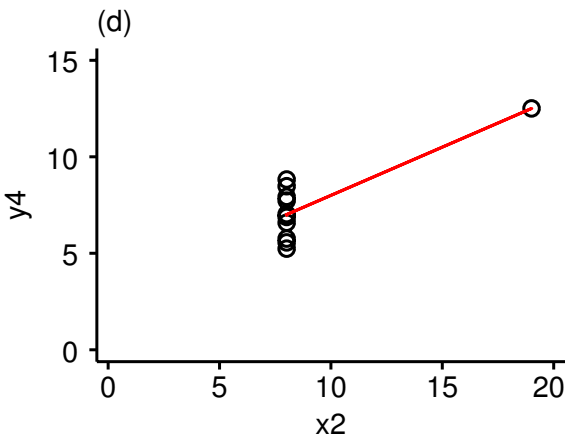


**Anscombe's quartett:** [Graphs in Statistical Analysis, Anscombe 1973, American Statistician]




$\hat{\beta}_0 = 3.0$
$\hat{\beta}_1 = 0.5$
$se(\hat{\beta}_1) = 0.118$
$R^2 = 0.67$
$\bar{X} = 9.0$
$S_X = 3.32$
$\bar{Y} = 7.5$
$S_Y = 2.03$

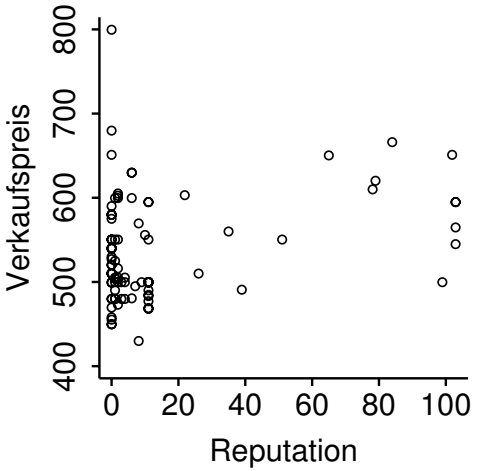
# **Reputation effects in internet auctions:** [Vertrauen und Reputationseffekte bei Internet-Auktionen, Diekmann und Wyder 2002, KZfSS]

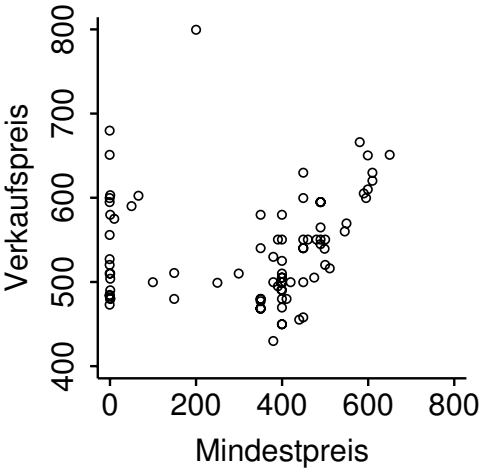
*Tabelle 3:* Effekt der Reputation auf den Verkaufspreis

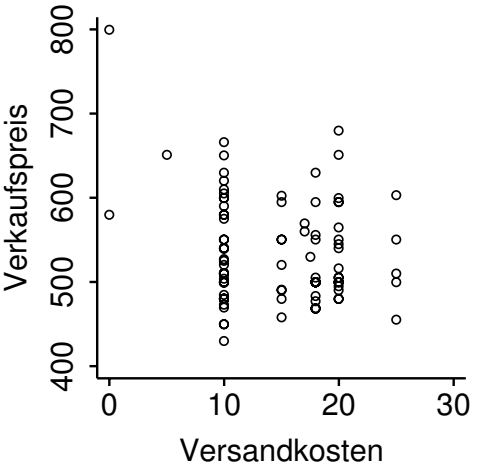
| | Modell 1 mit absoluter Anzahl Bewertungen | Modell 2 mit logarithmierter Anzahl Bewertungen | Modell 3 mit absoluter Anzahl Bewertungen und Heckman-Korrektur |
|---|---|---|---|
| Reputation (Anzahl Bewertungen) | 0,671** (3,19) | 10,755** (2,62) | 0,720*** (3,46) |
| Mindestpreis | 0,055 (1,19) | 0,075 (1,64) | –0,045 (–0,89) |
| Versandkosten | –2,549* (–2,48) | –3,111** (–2,82) | –2,048* (–2,03) |
| Dauer der Auktion | –0,200 (–0,16) | –0,569 (–0,45) | –0,080 (–0,067) |
| Mindesterhöhung | 3,313*** (4,29) | 3,635*** (4,64) | 2,923*** (4,05) |
| Anzahl der Gebote | 1,278 (1,89) | 1,597* (2,39) | 0,685 (1,02) |
| Konstante | 505,79*** (16,88) | 496,50*** (16,51) | 529,45*** (17,98) |
| Adj. $R^2$ | 0,261 | 0,237 | 0,408 |
| Lambda | – | – | 72,042*** (4,95) |
| N | 99 | 99 | 99 |

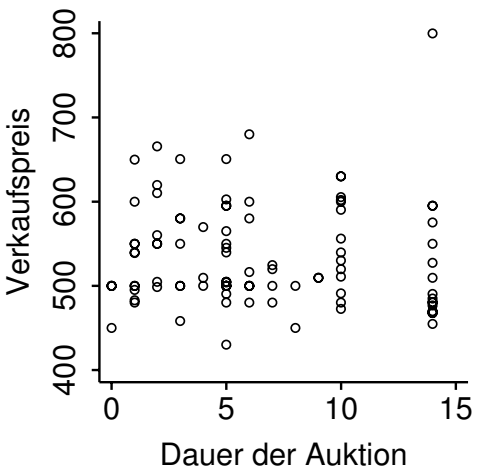
OLS-Regression mit der abhängigen Variable Verkaufspreis (netto, ohne Versandkosten), t-Werte in Klammern. Signifikant für: $\alpha = 0{,}05$ (*), $\alpha = 0{,}01$ (**), $\alpha = 0{,}001$ (***) bei zweiseitigem Test. Modell mit logarithmierter Reputation: Reputation = ln(Anzahl Bewertungen +1). Heckman-Korrektur: Probit-Schätzung des Verkaufserfolgs mit Prädiktoren „Reputation", „Mindestpreis" und „Versandkosten". Die geschätzte Korrelation zwischen dem Fehlerterm der Regression und der Probitgleichung ist größer als eins und wurde durch eins ersetzt.

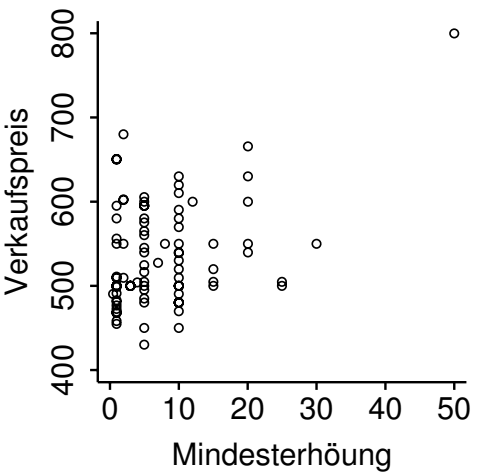
# Bivariate scatter plot:

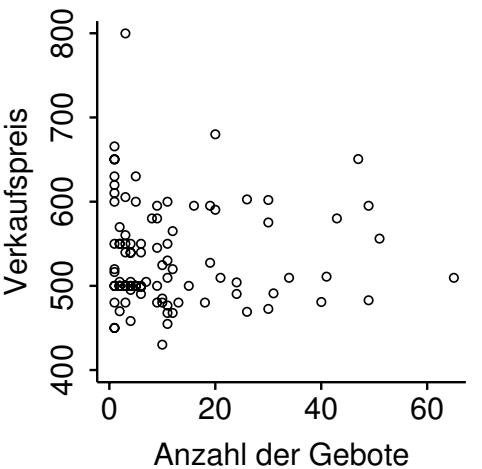


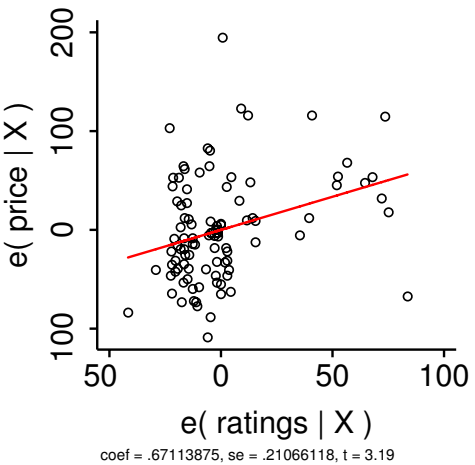



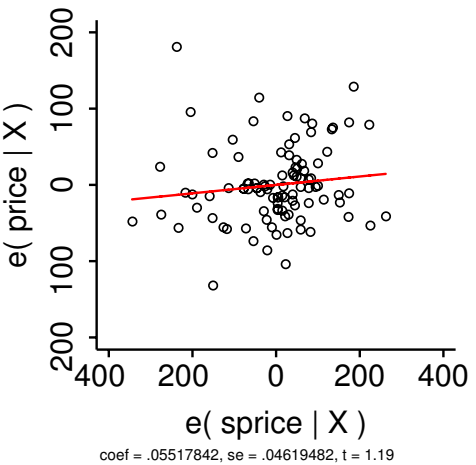



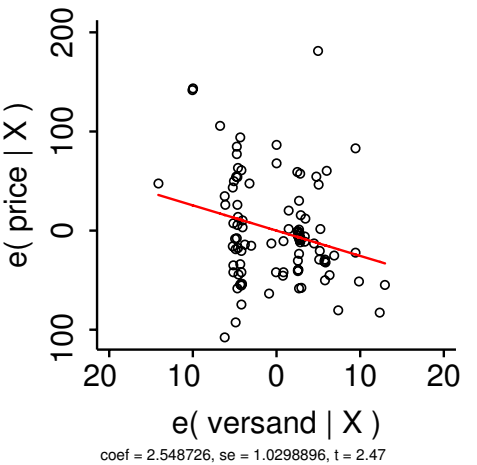



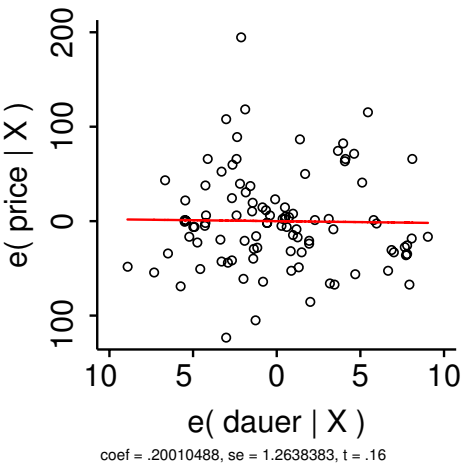



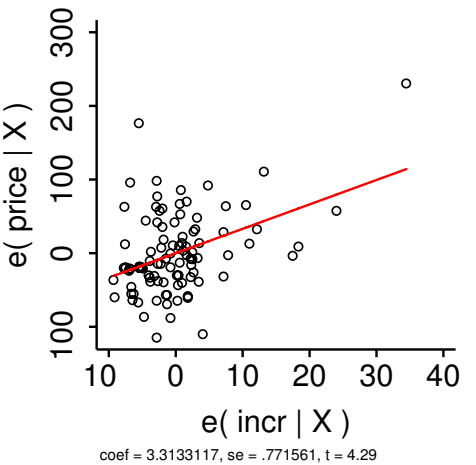
# Partial regression plot (added variable plot):

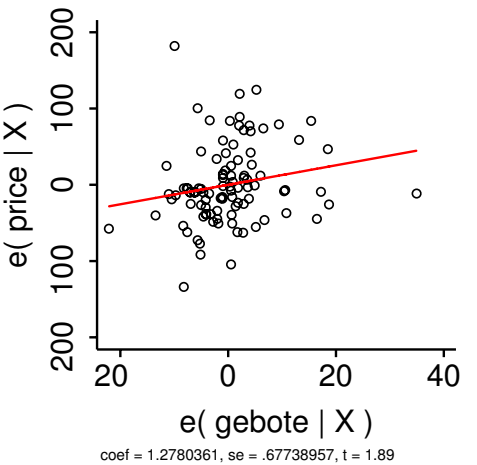


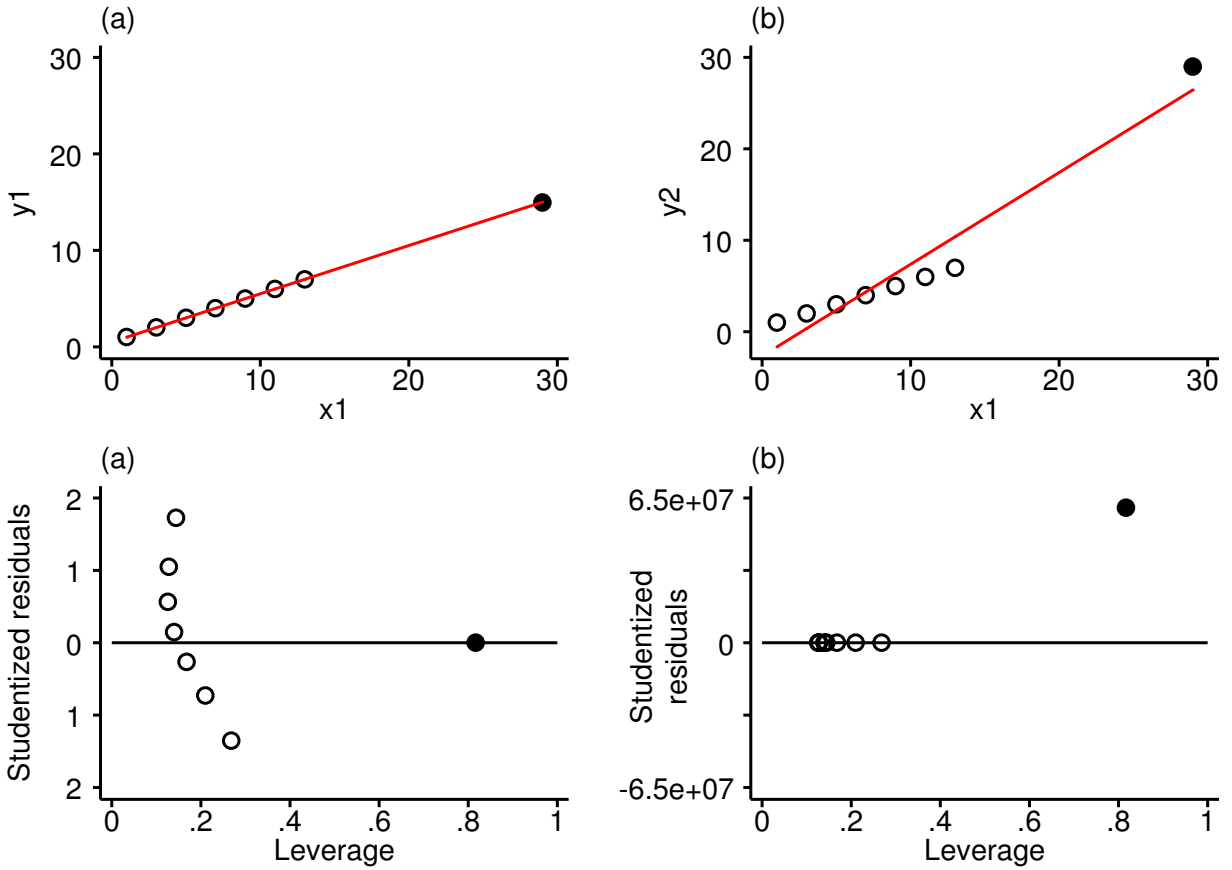



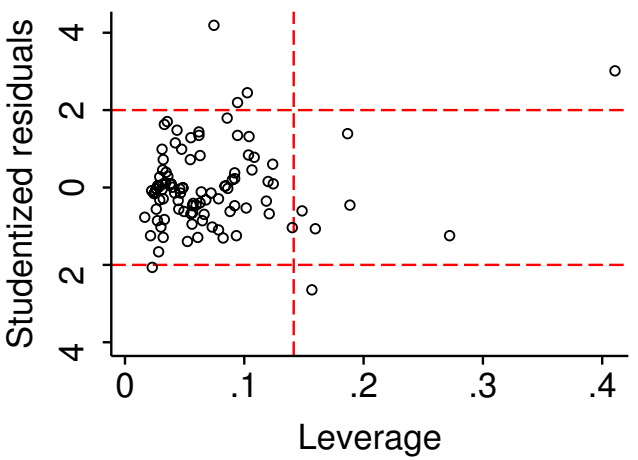



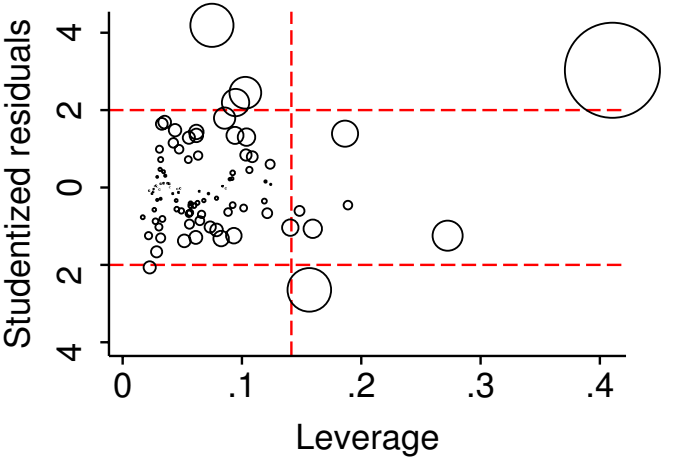



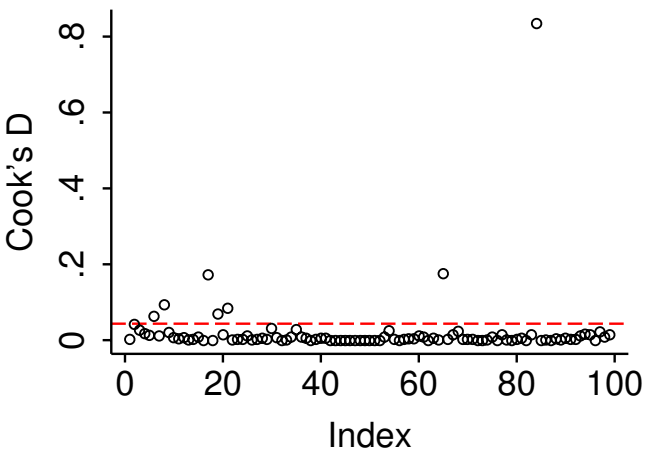
# Studentized residuals, leverage, Cook's D and DFFITS:

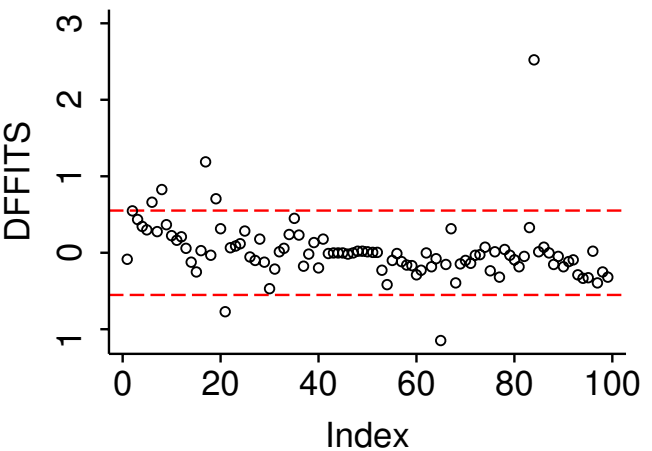


# Studentized residuals, leverage, Cook's D and DFFITS:

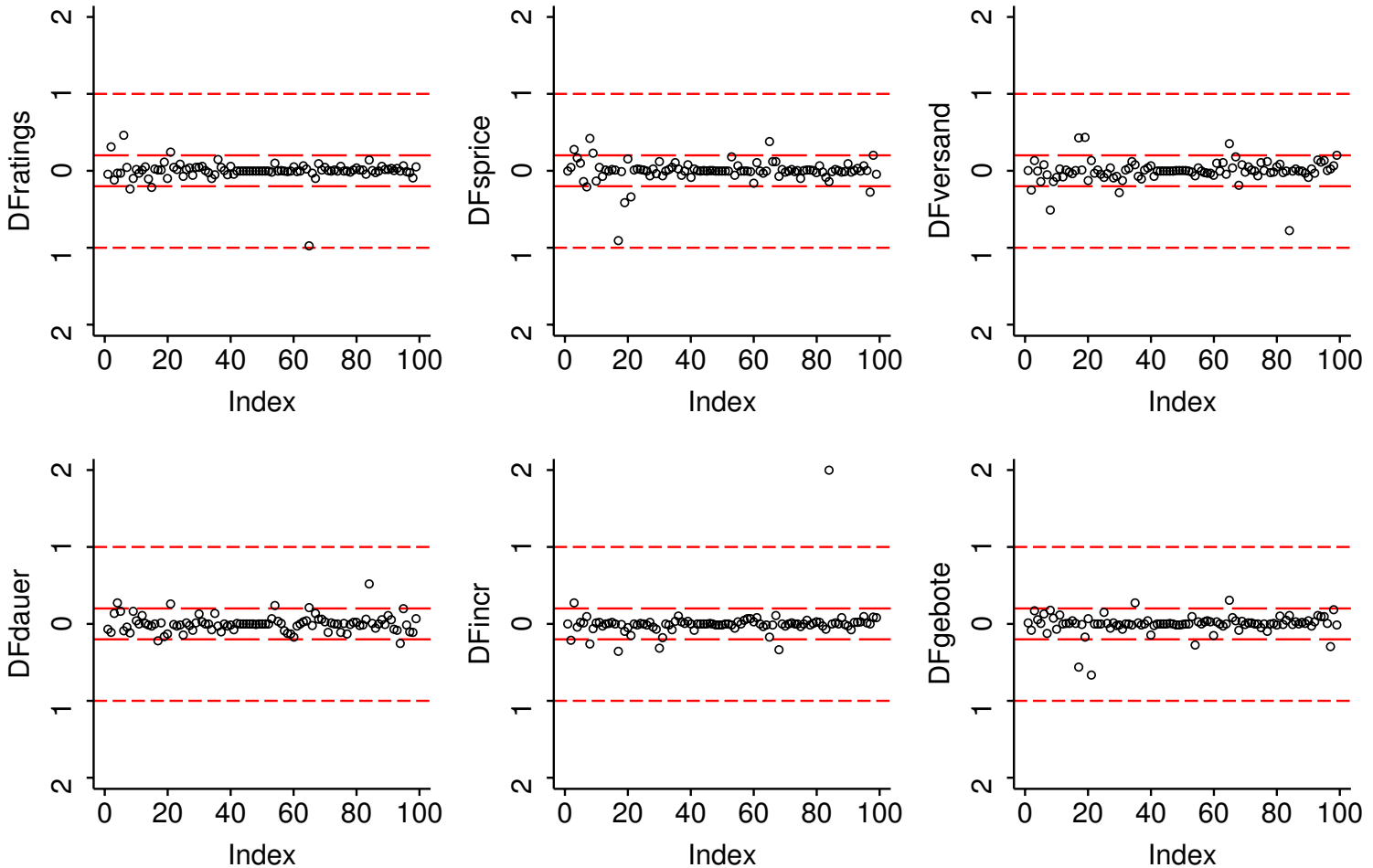


# DFBETAS:


DFratings
DFsprice
DFversand
DFdauer
DFincr
DFgebote
Index
2
1
0
1
2
0
20
40
60
80
100

# Outliers (DFFITS, Cook's D)

| ID | price | h | r | DFFITS | D |
|---|---|---|---|---|---|
| 6 | 651 | **0.187** | 1.391 | 0.666 | 0.063 |
| 8 | 651 | 0.103 | **2.445** | 0.827 | 0.093 |
| 17 | 680 | 0.075 | **4.189** | 1.192 | 0.172 |
| 19 | 603 | 0.094 | **2.197** | 0.710 | 0.069 |
| 21 | 510 | **0.272** | -1.255 | -0.767 | 0.084 |
| 65 | 500 | **0.156** | **-2.654** | -1.143 | 0.175 |
| 84 | 800 | **0.410** | **3.023** | 2.522 | 0.835 |

| ID | ratings | sprice | versand | dauer | incr | gebote |
|---|---|---|---|---|---|---|
| 6 | **0.458** | -0.140 | 0.075 | -0.089 | 0.013 | 0.131 |
| 8 | **-0.239** | **0.423** | **-0.509** | -0.118 | **-0.258** | 0.173 |
| 17 | 0.014 | **-0.911** | **0.427** | **-0.221** | **-0.353** | **-0.559** |
| 19 | 0.114 | **-0.417** | **0.431** | -0.168 | -0.096 | -0.170 |
| 21 | **0.247** | **-0.340** | 0.132 | **0.261** | -0.156 | **-0.664** |
| 65 | **-0.972** | **0.384** | **0.353** | **0.210** | -0.171 | **0.307** |
| 84 | 0.143 | -0.139 | **-0.778** | **0.517** | **1.999** | 0.111 |

Models without outliers:

| | all cases | | without no. 84 | | without outliers | |
|---|---|---|---|---|---|---|
| | coef | t | coef | t | coef | t |
| Reputation | 0.671 | 3.19 | 0.642 | 3.18 | 0.701 | 3.64 |
| Mindestpreis | 0.055 | 1.19 | 0.061 | 1.38 | 0.108 | 2.63 |
| Versandkosten | -2.549 | -2.47 | -1.781 | -1.75 | -2.897 | -3.22 |
| Dauer der Auktion | -0.200 | -0.16 | -0.826 | -0.67 | -0.456 | -0.44 |
| Mindesterhöung | 3.313 | 4.29 | 1.835 | 2.07 | 2.884 | 3.85 |
| Anzahl der Gebote | 1.278 | 1.89 | 1.206 | 1.86 | 1.726 | 2.77 |
| Constant | 505.8 | 16.88 | 506.4 | 17.63 | 487.7 | 19.82 |
| Adj. $R^2$ | 0.261 | | 0.164 | | 0.332 | |
| $n$ | 99 | | 98 | | 92 | |

Data reconsidered:

1. some minor coding errors
2. two doubly recorded cases
3. several inhomogeneous cases ("as good as new" instead of "new"; multiple offers; two for one)
4. "buy it now" feature neglected (right censoring)
5. information about accessoires neglected
6. variable "time" neglected
7. clustering on sellers neglected

# New models:

| | old model | | cleaned | | extended | |
|---|---|---|---|---|---|---|
| | coef | t | coef | t | coef | t |
| Reputation | 0.679 | 3.18 | 0.797 | 3.23 | 0.398 | 2.24 |
| Mindestpreis | 0.055 | 1.17 | 0.070 | 1.58 | 0.049 | 1.53 |
| Versandkosten | -2.469 | -2.37 | -2.674 | -2.66 | -1.988 | -2.80 |
| Dauer der Auktion | -0.097 | -0.08 | -1.161 | -0.94 | -2.792 | -3.08 |
| Mindesterhöung | 3.380 | 4.32 | 3.468 | 4.78 | 2.256 | 4.23 |
| Anzahl der Gebote | 1.293 | 1.89 | 1.698 | 2.64 | 0.992 | 2.15 |
| Time (centered) | | | | | -0.878 | -6.64 |
| Time (squared) | | | | | 0.011 | 3.20 |
| Accessoires | | | | | 36.988 | 5.59 |
| Constant | 503.0 | 16.52 | 499.1 | 17.35 | 511.2 | 24.65 |
| Adj. $R^2$ | 0.264 | | 0.368 | | 0.697 | |
| $n$ | 97 | | 84 | | 84 | |

# Taking into account left/right censoring and clustering on sellers:

```
Interval regression                               Number of obs   =        167
                                                  Wald chi2(9)    =     433.39
Log pseudo-likelihood = -318.79951                Prob > chi2     =     0.0000

                              (standard errors adjusted for clustering on aid)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ratings |   .5947603   .1288933     4.61   0.000     .3421341    .8473864
      spreis |  -.0149354   .0295831    -0.50   0.614    -.0729172    .0430465
     versand |    -1.9631   .8192902    -2.40   0.017     -3.56888   -.3573211
       dauer |  -4.343765   .8056098    -5.39   0.000    -5.922732   -2.764799
      erhoeh |   1.399547    .821364     1.70   0.088    -.2102964    3.009391
      gebote |   .3108419   .4106931     0.76   0.449    -.4941017    1.115785
      c_time |  -.7639231   .1292998    -5.91   0.000    -1.017346   -.5105002
     c_time2 |   .0121172   .0048581     2.49   0.013     .0025955    .0216389
    zubehoer |   32.01286   10.76621     2.97   0.003     10.91149    53.11424
       _cons |   563.0477   19.83601    28.39   0.000     524.1698    601.9255
-------------+----------------------------------------------------------------
    /lnsigma |   3.541923   .0967558    36.61   0.000     3.352285    3.731561
-------------+----------------------------------------------------------------
       sigma |   34.53327   3.341295                      28.56794    41.74422
------------------------------------------------------------------------------

  Observation summary:        59     uncensored observations
                              83  left-censored observations
                              25 right-censored observations
                               0       interval observations
```

```
Interval regression                               Number of obs   =        167
                                                  Wald chi2(9)    =     282.89
Log pseudo-likelihood = -318.33777                Prob > chi2     =     0.0000

                              (standard errors adjusted for clustering on aid)
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lnratings |   10.21581   2.954549     3.46   0.001     4.424998    16.00662
      spreis |  -.0104967   .0316813    -0.33   0.740     -.072591    .0515975
     versand |  -2.720376   .7964047    -3.42   0.001    -4.281301   -1.159452
       dauer |  -4.521466   .8267566    -5.47   0.000    -6.141879   -2.901053
      erhoeh |   1.611422   .7963867     2.02   0.043     .0505326    3.172311
      gebote |   .4399539    .436121     1.01   0.313    -.4148275    1.294735
      c_time |  -.7306907   .1301975    -5.61   0.000    -.9858731   -.4755083
     c_time2 |   .0118916   .0046313     2.57   0.010     .0028144    .0209688
    zubehoer |   30.81795   10.52142     2.93   0.003     10.19634    51.43956
       _cons |   562.2065   21.24511    26.46   0.000     520.5668    603.8461
-------------+----------------------------------------------------------------
    /lnsigma |   3.538714   .0951093    37.21   0.000     3.352303    3.725125
-------------+----------------------------------------------------------------
       sigma |   34.42262    3.27391                      28.56845     41.4764
------------------------------------------------------------------------------

  Observation summary:        59     uncensored observations
                              83  left-censored observations
                              25 right-censored observations
                               0       interval observations
```